# Régression logistique polytomique pénalisée à logits cumulatifs

Clémence Karmann [1,2] & Anne Gégout-Petit [1,2]

[1] *INRIA Nancy, équipe BIGS, 615 Rue du Jardin-Botanique, 54600 Villers-lès-Nancy ; clemence.karmann@inria.fr*

[2] *Université de Lorraine, Institut Elie Cartan de Lorraine, UMR 7502 ; anne.gegout-petit@univ-lorraine.fr*

**Résumé.** Notre but est d'étudier la dépendance entre une variable réponse prenant plusieurs modalités ordonnées et des variables explicatives quantitatives. C'est par exemple le cas si on étudie l'influence de certains facteurs sur différents stades du cancer ou sur différents niveaux de douleurs. Nous introduirons dans ce cadre le modèle de régression logistique à logits cumulatifs qui est une généralisation de la régression logistique, ainsi que l'estimation lasso des coefficients de régression grâce à l'algorithme de Frank-Wolfe. Enfin, nous présenterons quelques résultats de simulations.

**Mots-clés.** Régression logistique, polytomique, multi-classes, lasso, sélection de variables, algorithme de Frank-Wolfe

**Abstract.** Our goal is to study the dependance between a response variable and quantitative explanatory variables in the case where the response variable takes strictly more than two ordered modalities. For example, it happens when we study the influence of some factors on different stages of cancer or pain scales. In this framework, we introduce the polytomous logistic regression model using cumulative logits which generalizes the logistic regression. Then, we present the lasso estimation of the regression coefficients thanks to the Frank-Wolfe algorithm. Finally, we provide some experimental results.

**Keywords.** Logistic regression, polytomous, multiclass, lasso, variable selection, Frank-Wolfe algorithm

# 1 Introduction

Dans cet article, on présente une généralisation de la régression logistique pour une variable réponse prenant plus de deux modalités ordonnées. En effet, notre objectif est de modéliser la dépendance d'une certaine variable réponse sur des variables explicatives dans un cas où la variable réponse prend plusieurs modalités ordonnées. C'est par exemple le cas si notre variable réponse correspond aux différents stades d'un cancer, à différents niveaux de douleurs ou encore lorsqu'on peut regrouper en classes ordonnées les valeurs d'une variable. Plus particulièrement, on s'intéresse à mesurer l'influence de chacun des prédicteurs sur la variable réponse et donc à déterminer quels prédicteurs sont significatifs.

# 2 Régression polytomique ordonnée à logits cumulatifs

L'idée est d'étendre le modèle de régression logistique à une variable réponse qui prend plus de deux modalités (ordonnées). Ce modèle est décrit par Agresti (2010, 2013) ou encore Nakache et Confais (2003).

On considère $Y$ une variable quantitative qui prend $K$ modalités ordonnées et $p$ variables explicatives quantitatives $X_1, ..., X_p$. On modélise alors $\forall j \in \{1, ..., K-1\}$ :

$$\text{logit } \mathbb{P}_\beta(Y \leq j | X = x) = \text{logit} p^j_\beta(x) = \alpha_j + \beta_1 x_1 + ... + \beta_p x_p$$

c'est-à-dire,

$$\mathbb{P}_\beta(Y \leq j | X = x) = p^j_\beta(x) = \frac{\exp(\alpha_j + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\alpha_j + \beta_1 x_1 + ... + \beta_p x_p)}$$

Les coefficients $\beta$ ne dépendent pas du niveau $j$ de la variable réponse et comme $\mathbb{P}_\beta(Y \leq K | X = x) = 1$, il y a $p + K - 1$ coefficients à estimer.

De la même façon que pour la régression logistique, on peut considérer les odds de ce modèle pour mieux comprendre le rôle des coefficients. En particulier, on a que $\forall j \in \{1, ..., K-1\}$, $\dfrac{\text{odds}_j(x)}{\text{odds}_j(x_i^{+z})} = \exp(\beta_i z)$ en notant $x = (x_1, ..., x_p)$ et $x_i^{+z} := (x_1, ..., x_i + z, ..., x_p)$.

Notre objectif est de sélectionner les variables significatives du modèle, c'est-à-dire celles pour lesquelles le coefficient de régression $\beta_i$ (qui mesure également la dépendance conditionnelle) est non nul. C'est pourquoi on s'intéresse plus particulièrement aux coefficients $\beta$. Au lieu de tester la nullité de chaque coefficients $\beta_i$ (ce qui est assez chronophage), on va ajouter une pénalisation $L_1$ sur les coefficients $\beta$ lors de l'estimation des coefficients du modèle.

# 3 Estimation, inférence

Sans pénalisation, l'estimation des coefficients $\alpha$ et $\beta$ se fait par maximisation de la vraisemblance.

## 3.1 Optimisation lasso

On veut estimer les coefficients $\alpha$ et $\beta$ en pénalisant la log-vraisemblance $L$ sur les coefficients $\beta$. Pour cela, on a besoin de résoudre le problème d'optimisation suivant :

$$\underset{\substack{\alpha \in A \\ \beta \in \mathbb{R}^p}}{\text{argmax}} \left\{ L(\alpha, \beta) - \lambda ||\beta||_1 \right\}$$

qui est équivalent (par dualité lagrangienne) au problème :

$$\underset{\substack{\alpha\in A\\ \beta\in B_t}}{\operatorname{argmax}} L(\alpha,\beta) \tag{1}$$

où $\alpha = (\alpha_1, ..., \alpha_{K-1})$, $\beta = (\beta_1, ..., \beta_p)$ et $A$ et $B_t$ désignent les ensembles convexes suivants : $A := \{(\alpha_1, ..., \alpha_{K-1}) \in \mathbb{R}^{K-1} \,/\, \alpha_1 < ... < \alpha_{K-1}\}$, $B_t := \{(\beta_1, ..., \beta_p) \in \mathbb{R}^p \,/\, ||\beta||_1 \leq t\}$.

Nous avons résolu ce problème d'optimisation à l'aide de l'algorithme de Frank-Wolfe (Frank et Wolfe (1956)) qui est décrit plus en détails ci-dessous. On note $\beta^* = (\alpha_1, ..., \alpha_{K-1}, \beta_1, ..., \beta_p)$ et $X_j^{*i} = (0, ..., 1, ..., 0, X_1^i, ..., X_p^i)$ où 1 est en $j$-ème position.

Algorithme de Frank-Wolfe

1. On démarre avec une valeur initiale $\beta_0^* = (\alpha_0, \beta_0)$.
   À chaque itération $k$ :

2. On résout : $s_k \in \underset{\substack{s_\alpha\in A\\ s_\beta\in B_t}}{\operatorname{argmin}} {}^t(-\nabla L(\beta_k^*)) \begin{pmatrix} s_\alpha \\ s_\beta \end{pmatrix}$.

3. Le nouvel itéré est $\beta_{k+1}^* = (1-\gamma_k)\beta_k^* + \gamma_k s_k$, $\gamma_k = \frac{2}{k+1}$.

4. On continue jusqu'à convergence.

Dans notre cas, on peut séparer le problème d'optimisation de l'étape 2 en deux problèmes d'optimisation différents : $\alpha_k \in \underset{s\in A}{\operatorname{argmin}} {}^t(-\nabla L(\beta_k^*))_{|\alpha} s$ (optimisation linéaire sous contraintes) et $\beta_k \in \underset{s\in B_t}{\operatorname{argmin}} {}^t(-\nabla L(\beta_k^*))_{|\beta} s$.

L'optimisation relative aux $\beta$ est équivalente à résoudre $-t \underset{s\in B_1}{\operatorname{argmax}} {}^t(-\nabla L(\beta_k^*))_{|\beta} s$. Ceci revient à choisir $i_k$ tel que $\mid -\nabla_{i_k} L(\beta_k^*) \mid$ maximise la valeur absolue du gradient (la partie du gradient relative aux $\beta$ plus précisément). On obtient alors la solution : $-t \operatorname{sign}(-\nabla_{i_k} L(\beta_k^*))\, e_{i_k}$ où $(e_i)_{i=1}^p$ désigne la base canonique de $\mathbb{R}^p$.

## 3.2 Choix du paramètre

Nous avons utilisé deux méthodes pour la sélection des variables. En fait, ces deux méthodes ne conduisent pas au choix d'un paramètre $t$ (pour l'optimisation (1)) mais selectionnent directement les variables significatives. La première est la méthode de Stability selection proposée par Meinshausen et Bühlmann (2010) et la seconde est inspirée de la méthode des knockoffs de Candes (2015).

### 3.2.1 Stability selection

L'idée est d'estimer la probabilité qu'une variable soit dans le modèle afin de déterminer quelles variables sont les plus significatives.

On considère un ensemble $T$ de valeurs pour le paramètre $t$. Pour tout $t \in T$, on veut estimer la probabilité $p_i(t)$ pour chaque variable $i$ d'être dans le modèle. Pour cela, on effectue la régression pénalisée sur B ensembles d'observations obtenus par bootstrap. La probabilité estimée $\hat{p}_i(t)$ est alors la proportion de sélection de la variable $i$ parmi les B régressions bootstraps ($t$ est fixé).

Au lieu de choisir un modèle $\hat{S}^t$ défini par un certain $t \in T$, on choisit ici le modèle $\hat{S} := \{i \in \{1, ..., p\} \ : \ \max_{t \in T} \hat{p}_i(t) \geq p_{thr}\}$ pour un seuil fixé $p_{thr}$.

### 3.2.2 Knockoffs revisités

Le principe est d'utiliser une matrice $\tilde{X}$ de copies (des variables $X_i$) dont la structure est similaire à celle de $X$ mais non liée à $Y$. Le but est de déterminer si une variable est significative en étudiant si elle rentre dans le modèle avant ou après sa copie.

On construit notre matrice de copies $\tilde{X}$ en permutant les lignes (c.à.d. les individus) de la matrice design $X$. De cette façon, les corrélations entre les copies restent les mêmes. On note $\hat{\beta}(t)$ les coefficients estimés de la régression pénalisée de $Y$ sur la matrice augmentée $[X, \tilde{X}]$. Pour chaque variable $i$ de la matrice design augmentée, on considère $Z_i := \inf \{t > 0, \ \hat{\beta}_i(t) \neq 0\}$. Ceci nous fournit un $2p$-vecteur $(Z_1, ..., Z_p, \tilde{Z}_1, ..., \tilde{Z}_p)$. Ensuite, on considère, pour tout $i \in \{1, ..., p\}$, $W_i := Z_i \wedge \tilde{Z}_i$. $\begin{cases} +1 & \text{si } Z_j < \tilde{Z}_j \\ -1 & \text{si } Z_j \geq \tilde{Z}_j \end{cases}$.

Une valeur positive pour $W_i$ signifie donc que la variable $i$ est rentrée dans le modèle avant sa copie. On sélectionnera donc les plus petites valeurs positives du $p$-vecteur de statistiques $W$.

# 4 Simulations

Nous avons effectué plusieurs types de simulations dans différents contextes (différentes lois pour les $X$, plusieurs valeurs de $n$ et $p$, différentes valeurs des coefficients $\beta$ de la régression). Nous montrons ci-dessous à titre d'exemple les résultats issus des simulations suivantes :

- $n = 200$ observations et $p = 50$ variables
- La matrice design des variables explicatives $X$ est simulée à l'aide de variables gaussiennes indépendantes centrées (et réduites pour éviter le poids de la variance dans l'estimation des coefficients)

- Les coefficients de régression sont : $\beta = (1, 1, 1, 1, 0, \ldots, 0)$. Les quatre premières variables sont les seules significatives.
- Nous utilisons les méthodes de Stability selection et des Knockoffs revisités décrites précédemment : pour le Stability selection, nous avons choisi $T = \{0.1, 0.4, 0.7, ..., 3.7\}$, $B = 100$ et $p_{thr} = 0.8$ et nous conservons $\max_{t \in T} \hat{p}_i(t)$ pour toute variable $i$. Pour les knockoffs, nous attribuons à chaque variable son "rang d'apparition".

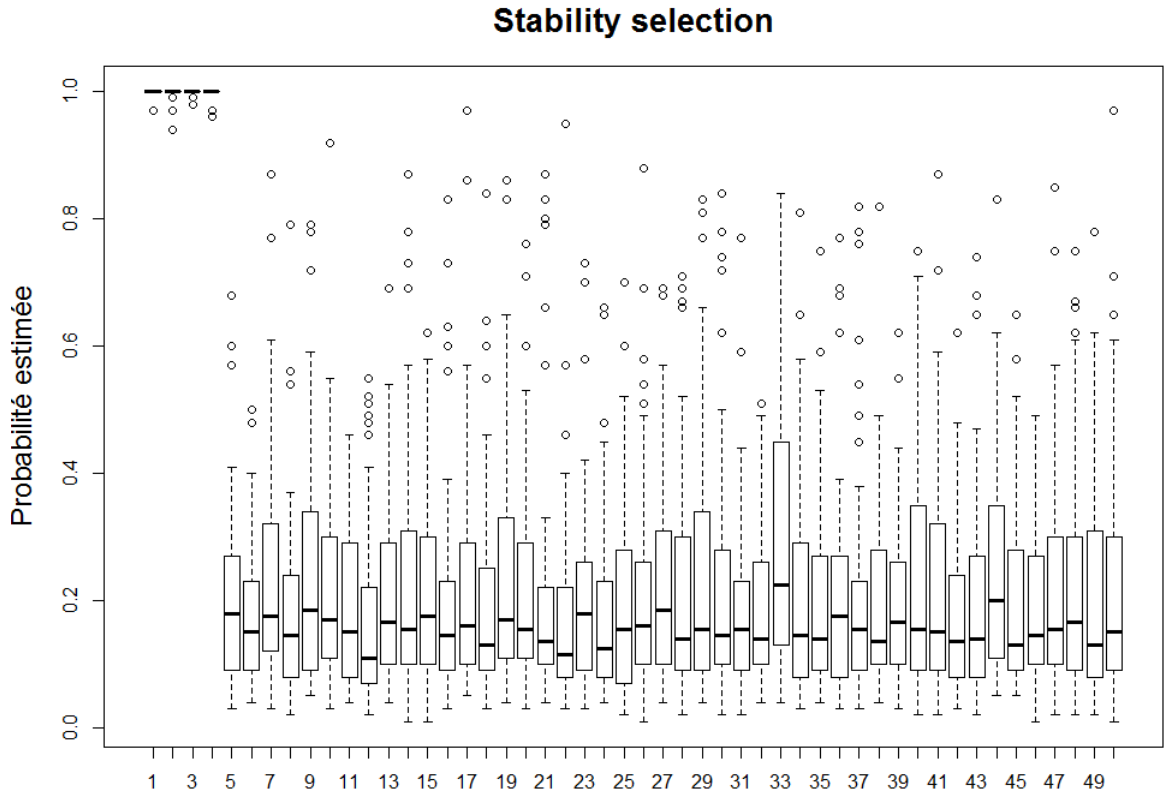


Figure 1: Boxplot sur 50 échantillons pour la méthode de Stability selection $-$ $n = 200, p = 50$

Concernant la méthode de Stability selection (figure 1), les quatre premières variables ont une probabilité estimée égale à 1 dans la quasi-totalité des cas alors que dans la moitié des cas, les autres ont une probabilité estimé inférieure à 0.3.

Les rangs d'apparition des variables 1 à 4 pour la méthode des knockoffs (figure 2) sont nettement inférieurs à ceux des autres variables. Ces rangs d'apparition correspondent à l'ordre dans lequel les variables rentrent dans le modèle, ainsi le rang 1 correspond à la première variable sélectionnée et ainsi de suite.

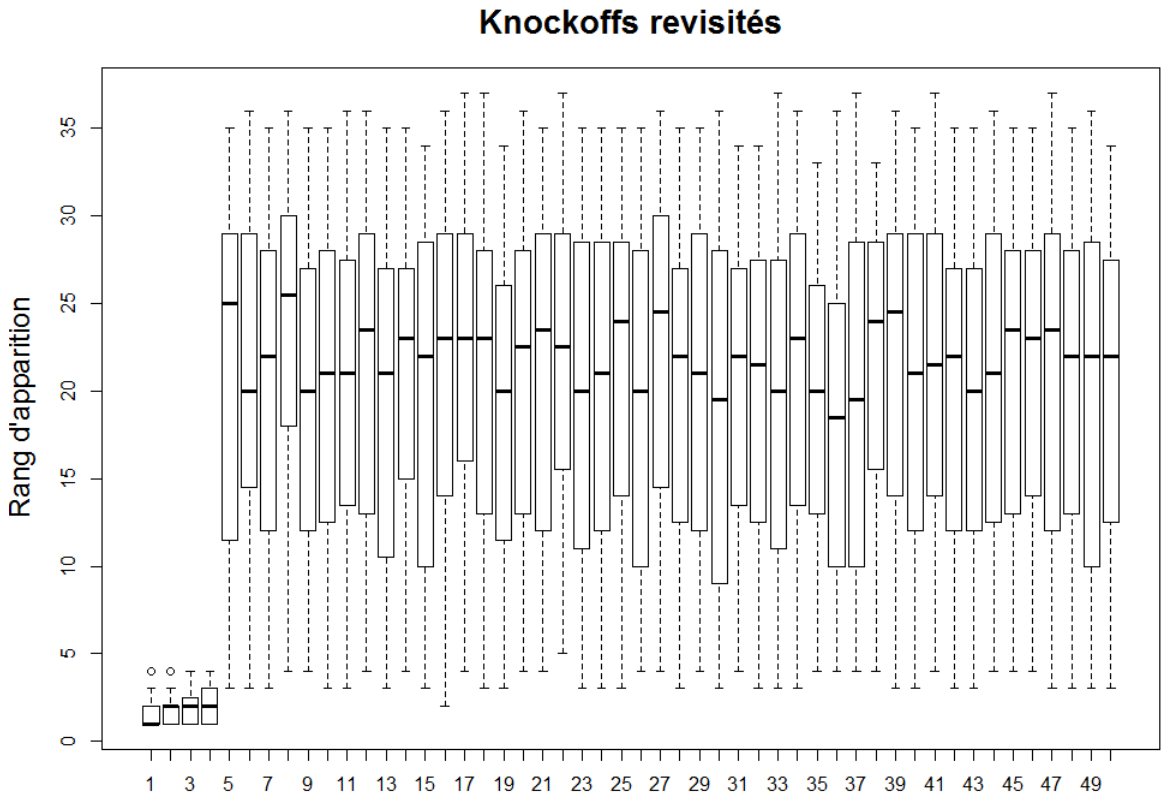


Figure 2: Boxplot sur 100 échantillons pour la méthode des knockoffs revisités $-$ $n = 200, p = 50$

# Bibliographie

[1] Agresti A. (2010), Analysis of ordinal categorical data, *John Wiley & Sons, Inc.*, New York.
[2] Agresti A. (2013), Categorical data analysis, *Wiley*, New York.
[3] Barber R. et Candès E. (2015), Controlling the false discovery via knockoffs, *Ann. Statist.*, 43, 2055–2085.
[4] Frank M. et Wolfe P. (1956), An algorithm for quadratic programming, *Naval. Res. Logist. Quart.*, 3, 95–110.
[5] Meinshausen N. et Bühlmann P. (2010), Stability selection, *J. R. Stat. Soc.*, 72, 417–473.
[6] Nakache J. et Confais J. (2003), Statistique explicative appliquée, *Editions Technip*, Paris.